DOCUMENT RESUME

ED 366 624 TM 020 999

AUTHOR Archbald, Doug A.

TITLE Authentic Assessment: What It Means and How It Can Help Schools.
INSTITUTION National Center for Effective Schools Research and Development, Madison, WI.
PUB DATE Dec 91
NOTE 34p.
AVAILABLE FROM National Center for Effective Schools Research and Development, 1025 W. Johnson Street, Suite 685, Madison, WI 53713 ($8).
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Context Effect; *Educational Assessment; Educational Background; Educational Innovation; Elementary Secondary Education; Incentives; Logical Thinking; Mastery Learning; Rote Learning; Scores; *Standardized Tests; Standards; *Student Evaluation; Test Interpretation; *Test Use
IDENTIFIERS *Authentic Assessment; *Performance Based Evaluation

ABSTRACT

Recent years have seen a new and serious commitment to improving methods of assessing academic performance. Schools, school districts, and states are experimenting with a wide range of assessment alternatives. This paper is about this new commitment to assessment and begins with some background on standardized tests because the rationale for authentic assessment requires understanding the purposes and limitations of standardized tests. It is argued that an authentic test is one that measures performance on authentic intellectual tasks in a way that reveals precisely what a student knows or can do. The second part of the paper discusses the potential of authentic assessment to improve curriculum and instruction. The following potentials are noted: (1) clarifying and legitimating high standards; (2) providing incentives for authentic instruction and tasks; and (3) providing more useful instructional and management information. It must be recognized that authentic assessment is costly and time-consuming, but its return is recognized in the fulfillment of the potentials described. Five figures illustrate the discussion. (Contains 41 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

AUTHENTIC ASSESSMENT: WHAT IT MEANS AND HOW IT CAN HELP SCHOOLS

Doug A. Archbald
University of Delaware

December, 1991

This occasional paper was prepared for the National Center for Effective Schools Research and Development, Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison, which is supported by the Spencer T. and Ann W. Olin Foundation. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the supporting agencies.

I wish to thank Jim Keefe, Monte Moses, and Fred Newmann for their helpful ideas in improving this paper.

© 1991 by the Board of Regents of the University of Wisconsin-Madison. All rights reserved.

INTRODUCTION

No one knows precisely how many standardized tests are taken each year by students. The National Center for Fair and Open Testing estimates K - 12 students in the U.S. take around 100 million standardized tests, about an average of 2.5 standardized tests per student per year (including state and district testing). The National Commission on Testing and Public Policy estimates standardized testing costs between $700 million to $900 million yearly in purchasing costs and administration time, or about $17 to $22 per student per year.

Is all this standardized testing necessary? Is it is making schools better? Proponents believe standardized tests help teachers and administrators make more informed decisions about instruction and programs as well as provide policymakers and the public with information about student achievement and school performance.

Others are skeptical and argue we should reduce (or stop) standardized testing.[1] However, most agree that standardized, multiple-choice tests have definite limitations, are overused and over-interpreted, and are unlikely to help schools achieve the reform goals envisioned in such prestigious reports as the Curriculum and Evaluation Standards of the National Council of Teachers of Mathematics, Everybody Counts, Science for All Americans, and Charting a Course: Social Studies for the 21st Century.

Recent years have seen a new and serious commitment to improving methods of assessing academic performance. Schools, districts, and states throughout the nation are experimenting with a wide range of assessment alternatives.

This paper is about this new commitment to assessment. It begins with some background on standardized tests because the rationale for authentic assessment requires understanding the purposes and limitations of standardized tests. Then it argues that an authentic test is one that measures performance on authentic intellectual tasks in a way that reveals precisely what a student knows or can do. The second part of this paper discusses the potential of authentic assessment to improve curriculum and instruction. It concludes with comments about feasibility.

SOME BACKGROUND ON STANDARDIZED TESTS

Experts disagree whether standardized tests can measure "higher order" thinking precisely. However, it is well documented that standardized tests measure some form of academic ability. For instance, people differ in "test-taking ability," at least with respect to standardized tests of general achievement and ability. Under normal circumstances, a student who performs well on one test of academic ability or achievement will perform similarly well on most other standardized tests of academic achievement or ability.[2] Moreover, a person's performance (beyond very young ages) on these types of tests remains relatively constant over the years.

The abilities and knowledge that contribute to test performance also contribute to academic performance in the classroom. Scores on standardized tests correlate with grades from about .5 to about .8. Correlations vary, depending on such factors as how close in time the measures of test score and grade point average are taken and the degree of match between the test's content and the curriculum on which a student has been graded. Most studies show the correlation between high school senior year standardized test scores and first-year college grades to be about .35 (Linn, 1982). This lower correlation reflects several factors: the year or so lapse in time between the two measures (and the fact that time brings change in people); the fact that the college GPA reflects a diversity of courses, some of which may be very different in cognitive demands compared with the standardized test; and the fact that there is more academic competition at the college level. (For statistical reasons, this last factor results in correlations lower than they would be otherwise if all high school test-takers went to college and produced a GPA to go into the sample for comparison purposes.)

How strong is the relationship denoted by correlations of .5 or .8? A .5 correlation means there is about a two-thirds probability that a student in the middle of the test score distribution will score somewhere between the 20th and the 80th percentile on the GPA distribution. (Conversely, the chances are one in three that the student might score above the 80th percentile or below the 20th percentile on the GPA distribution.) At a correlation of .35, the chances are one in three that a

student at the 90th percentile test score will place in the top fifth of the freshman class on grade point average (and two out of three that s/he will not).

Despite these probabilities, it is clear that standardized tests measure knowledge and abilities that contribute to academic achievement. For instance, higher scores on vocabulary items like the one in Figure 1 are evidence of a larger vocabulary, which is related to linguistic competence.

Choose the word opposite in meaning to the underlined word: timorous

(a) succinct **(b)** chaste
(c) audacious **(d)** insouciant

Figure 1: Sample Vocabulary Multiple-Choice Item

And higher scores on the kinds of problems in Figure 2 are evidence of greater mathematical knowledge. Other kinds of items can measure certain forms of logical skills. Critics of standardized tests sometimes neglect to acknowledge these reliable measurement properties of standardized tests, and that, while they are imperfect instruments, these tests efficiently provide information about student standing relative to peers on the domains tested.

1. Which of the following is equal to m^3?

   a) 3m **b)** $\sqrt[3]{m}$ **c)** m + 3 **d)** (m)(m)(m) **e)** none of the above

2. Pedro bought a record player at 20% off the list price. The list price was $95. How much did the record player cost?

   a) $70 **b)** $75 **c)** $76 **d)** $77 **e)** none of the above

Figure 2: Sample Mathematics Multiple-Choice Items

However, research is quite clear that standardized tests have serious limitations for assessing performance on open-ended problem-solving tasks -- tasks involving creativity, sustained inquiry, and analysis of diffuse problems (Baxter, Shavelson, Pine, & Yure, 1991; Frederiksen & Ward, 1978; Mentkowski & Doherty, 1982; Tindal & Parker, 1989; Ward, Frederiksen, & Carlson, 1980).[3]

CHARACTERISTICS OF AUTHENTIC INTELLECTUAL ACHIEVEMENT

What is the chief measurement problem with standardized, multiple-choice tests? The basic problem is that the cognitive demands of much academic work in school and intellectual work outside of school are quite different from the cognitive requirements of answering multiple-choice questions under standardized test conditions. Consider the cognitive challenges posed by the following tasks: deliberating over career or college choices; writing an essay on the meaning of political freedom; analyzing the logic of an editorial argument in a newspaper; critically evaluating the content of a politician's or an advertiser's rhetoric; collaborating on the design and content of a school newspaper, research project, or advertising brochure; or investigating ways to reduce shoplifting, drug abuse, teenage pregnancy. Competently performing these kinds of tasks requires knowledge, creativity, and problem solving.

We expect education to improve students' competence on these kinds of tasks, yet the competencies they require are quite unlike those measured by standardized, multiple-choice questions. There is a gap between educational purposes and measurement. Tests should assess performance on tasks that as closely as possible reflect authentic cognitive challenges and contexts. Yet our predominant and most visible form of testing is based on tasks that appear inauthentic -- that is, artificial or contrived -- and are widely distrusted. We have discussed this elsewhere as the problem of authenticity in assessment (Archbald & Newmann, 1988; Newmann & Archbald, in press).

The next section discusses, in more detail, the mismatch between authentic educational purposes and multiple-choice measurement. It describes key cognitive and situational

characteristics of authentic intellectual achievement and asks how well standardized tests are likely to measure these characteristics.

Cognitive Characteristics of Authentic Achievement

Producing Instead of Reproducing Ideas

The conceptual distinction between "producing ideas" and "reproducing ideas" is useful in thinking about the problem of assessing authentic achievement. Much student work in school involves *reproducing* ideas the student has memorized, such as definitions of terms, historical facts and dates, or mathematical formulas. The answer, "to make sugar from sunlight," to a question about photosynthesis may be no more than recalled information, and involve no creative process, and no idea generation. While authentic intellectual achievement requires a stock of remembered information, it also requires the *production* of new ideas. When a student develops a written argument or contemplates a complicated mathematics problem, then s/he draws not only on remembered information, but also creates new ideas.

Disciplined Production of Ideas

Ideas, of course, are not produced in the abstract. They are expressed in particular contexts and forms, and for particular purposes. Authentic intellectual achievement involves the production of ideas in spoken discourse, writing, artistic performances, designs, and material objects. But what makes ideas produced in these forms *more* as opposed to *less* authentic? Three features are paramount.

Background Knowledge

This refers to the need to know the related facts, concepts, and theories that together make one informed or literate about a particular area, whether it is algebra, computer programming, plant breeding, or early African civilization. While there are no clear boundaries on the amount of background knowledge needed to engage in authentic intellectual activity in a particular area, before a student can think analytically, critically, or creatively about particular issues or problems, s/he must have a base of knowledge on which to draw.

In-depth Understanding

Background knowledge provides the foundation to identify and investigate particular problems or issues. Depth of understanding grows from applying knowledge to new problems. Authentic cognitive work entails applying substantive and procedural knowledge to probe novel questions, gather new data, and create new explanations or solutions. A student with knowledge of ecological concepts such as energy transfer, food chains, and species interdependence, can use this knowledge to identify interesting questions and to conduct investigations. This leads to the creation of new knowledge.

Learning that is based mainly on reproducing memorized information lacks these qualities of authenticity. Reproducing memorized information may appear to involve the production of ideas, but it may be devoid of authentic intellectual work because the thinking involved is not disciplined by a prior knowledge base and does not deepen understanding in a particular area.

Integrating Ideas

Authentic intellectual achievement commonly requires integrating ideas. Developing a report, a talk, or a science investigation requires creating ideas and linking them to achieve particular purposes. The logic shaping this integrative process derives from the purposes to be achieved and makes the product a meaningful whole. This *wholeness* can be contrasted with the more *fragmented* nature of the many different kinds of repetitive drill and practice exercises in classrooms: worksheets on spelling, vocabulary, grammar, and computation; read-and-memorize assignments covering historical, geographic, or scientific facts. Such tasks mainly require students to *reproduce* facts, definitions, and formulas produced by others and memorized by the student. When students integrate ideas into academic products or performances, they are doing creative intellectual work.

Conventional standardized tests, and most multiple-choice tests for that matter, are well-suited to assess how well students can recall or recognize facts, definitions, and formulas. This is often useful information, but such tests are also limited in what they reveal. Multiple-choice tests

are unlikely to reveal whether a student has mastered and can apply a particular knowledge base or can integrate knowledge and skills from different areas.

Situational Characteristics of Authentic Achievement

Authentic achievement is not only characterized by certain cognitive requirements, but there are situational conditions characteristic of authentic intellectual achievement. Because a person's performance is highly influenced by *contextual factors*, it is useful to know what kinds of situational conditions would be worthwhile to reproduce in assessment situations. Four characteristics stand out.

First, and perhaps most important, authentic intellectual products and performances have *aesthetic and utilitarian purposes* that have value beyond evaluation. For example, a letter of complaint intends to elicit a desired action; a story aims to evoke empathy, sympathy, or other emotions to entertain and enlighten. People write to *communicate*, not just to show how *well* they communicate, just as they calculate to compute gasoline mileage or costs of retail items, not to display arithmetic skills, but to achieve particular purposes.

Second, authentic intellectual work *takes time*. Writing and revising, thinking and rethinking, trial and error -- the creative and analytical processes of authentic academic work are not easily pressed into short time slots. Time limits for authentic tasks must arise from the nature of the task -- a writing task may require 45 minutes, a speech several days, or a research task several weeks.

Third, authentic intellectual work is *often collaborative*. Students and assembly line workers[4] are about the only people who routinely are forced to work independently, under conditions where collaboration is explicitly prevented. Yet in offices, laboratories, and agencies everywhere people jointly produce ideas and solve problems, react to each other's work, and share advice.

Fourth, authentic intellectual work *often involves materials and equipment*. Today tools and technology pervade virtually all aspects of our lives. Calculators, electronic information

networks, telecommunications tools, computerized data-bases, statistical and graphics software, mechanical tools, measurement instruments, reference catalogues, and many other things are becoming integral fixtures in the production of information, solutions, and services.

It is clear that conventional, standardized, multiple-choice tests cannot capture intellectual performance in the context of all these conditions. Students perceive little intrinsic aesthetic or utilitarian value in taking standardized tests other than their sense of having their "ability" evaluated. Instead of allowing time for sustained thought, multiple-choice testing is conducted in short, intense bursts of timed activity. Instead of working collaboratively with functional materials, students take standardized tests by themselves with only a pencil; "collaboration" can bring charges of cheating.

Characteristics of Authentic Intellectual Achievement: Concluding Remarks

The central question addressed above is: What are key cognitive and contextual characteristics that make performance or achievement *authentic*? The discussion articulates a conception of competence responding to this question and argues standardized, multiple-choice questions are the wrong instrument to measure authentic intellectual performance.

Before turning to some examples of ways to assess authentic achievement, two qualifications are important. First, some claim that criticizing standardized tests for not providing direct measures of authentic intellectual performance is like claiming a barometer does not give a complete picture of the weather. Standardized tests are designed for limited purposes, primarily to rank students on specific skills related to academic ability (Anastasi, 1988). Tests of general achievement or aptitude are not designed to evaluate curriculum programs, particular instructional techniques, students' depth of knowledge in particular areas, or vague attributes like "intelligence." However, what standardized tests are specifically designed to do is less the issue than how they are used. Because test scores are often over-interpreted and are often used to make evaluative decisions about whether students, teachers, programs, and schools are achieving broad educational

purposes, it is fair and indeed necessary to stress the significant limitations of standardized tests in relation to the many purposes for which they are used.

Second, it should be understood that "authentic achievement" as described above is a set of ideal principles. Qualities of human accomplishment can be called authentic when they are rooted in genuine human purposes and endeavors that challenge and engage the intellect. True, it is unlikely that assessment done in schools and classrooms will always reflect principles of authenticity. Still these principles are useful as guides to improve assessment practice and to sharpen questions about validity in multiple-choice measurement.

CHARACTERISTICS OF AUTHENTIC ASSESSMENT

A rich variety of assessment criteria, methods, and tasks go under the heading "assessing authentic achievement." Before presenting some examples, it is useful to begin by describing three features common to all types of authentic assessment.

Multiple Criteria

Authentic intellectual tasks tend to be open-ended, requiring problem solving and creative thinking. One must produce ideas and draw upon multiple skills and bases of knowledge. Thus, the quality of authentic achievement should be *assessed on multiple criteria.* For example, foreign language proficiency might be judged on vocabulary, syntax, and accent; a science project might be evaluated on use of evidence, presentation of results, logic of conclusions, and clarity in communication.

The analogue of multiple criteria in a standardized test is the individual score disaggregated by subskill scores. While useful for certain purposes, the multiple-choice format limits how genuinely distinct the subskills are compared with the more diverse dimensions of knowledge and competence underlying authentic production tasks.

Prespecified Standards To Judge Quality

Each of the criteria on which a particular product or performance is rated must have *clearly written (or understood) standards to permit judgments* of "better or worse." These standards

explain differences between higher and lower scores on a particular dimension. Figure 3 (p. 12) shows criteria (the rows) and standards (the columns) used in one school district's writing assessment system.

A student's performance on a standardized test, in contrast, is norm-referenced and is expressed typically as a percentile or grade equivalent score. This reveals a given student's rank relative to other students, but does not r*eveal precisely what the student knows or can do.* Reporting raw subscores on standardized tests for discrete topics or skills approximates *criterion referencing* because performance is measured in relation to specified objectives, but the decontextualized nature of these topics and skills make direct interpretation difficult, and so norm-referenced judgments are far more common.

Human Judgment

Machines can grade responses on multiple-choice questions, but they cannot determine whether an essay is well-written, a performance well-executed, or a solution well-devised. The application of human judgment is part of authentic assessment, just as it is part of hiring and promotion evaluations, book and movie reviews, decisions about grants and awards, and innumerable other formal and informal judgments.

EXAMPLES OF AUTHENTIC ASSESSMENT

A large variety of methods of authentic assessment are currently proposed and practiced. While exceptions can be found, most fall into one of the following categories: discrete competencies, exhibitions of achievement, and portfolios.

Assessment of Discrete Competencies

Assessment of discrete academic competencies refers to measures of relatively specific domains of performance under standard conditions. For example, writing, speaking and listening, reading, analysis, and problem-solving can be assessed by engaging students in specific tasks and

rating their performance according to prespecified standards, such as those shown for writing in Figure 3 (p. 12). These measurement techniques yield quantitative information that can be used to describe the performance of individual students relative to other students or to describe individual students' growth.[5] Following are examples of assessment methods that produce quantitative measures and that meet several criteria for authenticity in the areas of language and mathematical problem solving. These procedures can also be used to assess mastery of specific curriculum content.

Figure 4 (p. 14) is a sample item from the 1987 National Assessment of Educational Progress's Pilot Study of Higher-Order Thinking Skills. Students are required in this paper-and-pencil task to examine data on five children competing in three athletic events (i.e., frisbee toss, weight lift, and 50-yard dash), devise their own approach for computing and interpreting the data, decide which of the five children would be the overall winner, and explain why they selected a particular "winner."

Oral communication can be assessed based on models developed by the Assessment of Performance Unit (APU) of the Department of Education and Science in Northern Ireland.[6] This agency has developed a variety of methods to assess oral communication in five categories:

1. Instructing/directing;
2. Giving and interpreting information;
3. Narrating;
4. Describing and specifying; and
5. Discussion.

Specific assessment tasks are designed to be as realistic as possible in order to elicit the type of language students would be expected to use outside the test situation. In each task a student is asked to talk with another person or small group to achieve a particular purpose. The requirement for a realistic and relaxed setting for the students is balanced against the need for standardization which entails a pre-scripted instructional protocol.

	1	2	3	4	5	
O	Little or nothing is written Essay is disorganized and poorly developed. Does not stay on topic.		Essay is incomplete.It lacks an intro, well-developed body, or conclusion. Coherence and logic are attempted but inadequate		The essay is well-organized. It is coherent, ordered logically, and fully developed.	X 6
S	The student writes frequent run-ons or fragments.		Occasional errors in sentence structure. Little variety in sentence length structure		Sentences are complete and varied in length and structure	X 5
U	Student makes frequent error in word choice and agreement.		Student makes occasional errors in word choice and agreement.		Usage is correct. Word choice is appropriate.	X 4
M	Student makes frequent errors in spelling, punctuation and capitalization.		Student makes an occasional error in mechanics.		Spelling, capitalization, and punctuation are correct	X 4
F	Format is sloppy. There are no margins or indents. Handwriting is inconsistent.		The margins and indents have inconsistencies - no title or inappropriate title.		The format is correct. The title is appropriate. The margins and indents are consistent.	X 1

O = Organization S = Sentence Structure U = Usage M = Mechanics F = Format	The essays are scored using a 1 - 5 scale. The numbers in the boxes to the far right indicate the relative importance of each factor in the overall grade. Thus, organization is valued most, and counts 30% of the grade; format counts 5% of the grade.

Figure 3: Dimensions and Criteria for Analytical Scoring of Writing

Each student's performance is scored in three ways. During the oral task, students are given a holistic score (1 - 7), as well as an "orientation to the listener" (eye contact, non-verbal gestures) score (1 - 5). The assessor tapes each student's oral task, which is also evaluated later by a pair of trained assessors using analytical scoring techniques. For the analytical scoring, the categories are:

1. Sequential structure (organization). [1 - 5]
2. Lexico-grammatical features (lexical selection and syntax). [1 - 5]
3. Performance features (hesitancy/fluency, tempo/pacing, and verbal assertiveness). [1 - 3]

The Connecticut State Department of Education has developed a variety of assessment tasks, including the following, to assess "hands-on" investigations in science. These tasks are called "survival" investigations and require setting up and drawing conclusions from an experiment.

The student is presented with apparatus and the problem stated below. The student is given paper and a pencil; containers, lids, and two fabrics of different materials and sizes; a variety of construction materials (tape, string, etc.); thermometers, a ruler, and a watch; and paper towels. An assessor gives prescribed instructions and rates the student's performance using a scoring sheet.

Statement of the Problem: Imagine you are stranded on a mountainside in cold, dry, windy weather. You can choose a jacket made from one of the two fabrics in front of you. This is what you have to find out: Which fabric would keep you warmer?

The assessor may give the student suggestions for guidance -- to use a container filled with warm water (instead of a person), to make "jackets" for the container from the materials, to give each fabric a "fair test," and to clearly record all the steps and conclusions "so that someone else can understand what you have found out." Selections from the scoring guide for assessing students' performance on the "survival" task are shown in Figure 5 (p. 15).

Joe, Sarah, Jose, Zabi, and Kim decided to hold their own Olympics after watching the Olympics on TV. They needed to decide what events to have at their Olympics. Joe and Jose wanted a weight lift and a frisbee toss event. Sarah, Zabi, and Kim thought running a race would be fun. The children decided to have all three events. They also decided to make each event equally important. They held their Olympics one day after school. The children's parents were the judges and kept the children's scores on each of the events.

The children's scores for each of the events are listed below:

Child's Name	Frisbee Toss	Weight Lift	50-Yard Dash
Joe	40 yards	205 pounds	9.5 seconds
Jose	30 yards	170 pounds	8.0 seconds
Kim	45 yards	130 pounds	9.0 seconds
Sarah	28 yards	120 pounds	7.6 seconds
Zabi	48 yards	140 pounds	8.3 seconds

(A) Who would be the overall winner?

(B) Explain how you decided who would be the overall winner. Be sure to show all your work.

Rating Criteria for NAEP "Student Olympics" Item

4 - accurate ranking of children on each event; citing Zabi as overall winner.

3 - using a ranking approach, but misinterpreting performance on the dash event and naming incorrect winner.

2 - cites overall winner or a tie, with explanation that demonstrates recognition that a quantitative means of comparison is needed.

1 - selection of overall winner with a irrelevant, non-quantitative, or no explanation.

0 - no response

Figure 4: Item From NAEP Pilot: Assessing Higher-Order Thinking

Student Familiarity with Activity

1 = has done almost exact same exercise
2 = has done similar exercise
3 = has not done same or similar exercise

Approach Used (If student changed approach, encode the approach used that led to final conclusion.)

1 = fabric wrapped around hand - qualitative
2 = fabric around thermometer - qualitative
3 = fabric around thermometer - set temperature drop: time interval measured

Control - Fabric (size and fastening)

1 = controlled
2 = not controlled

Temperature Measurements

1 = all measurements within 2 degrees of test administrator's readings
2 = all except one or two measurements within 2 degrees of test administrator's readings
3 = more than 2 measurements different from test administrator's by more than 2 degrees

Recording of Data

1 = data organized and recorded clearly enough to permit appropriate interpretation
2 = data not organized and recorded clearly enough

Conclusion

1 = conclusion consistent with data
2 = conclusion not consistent with data
3 = conclusion not possible because of design or execution

Overall Evaluation of Experiment

1 = design and execution such that one could "trust" conclusion
2 = design and execution have minor problems which could create some doubt about conclusion
3 = design and execution such that one should have no faith in the conclusion at all

Figure 5: Selections from "Survival Experiment Scoring Guide" in Connecticut's Common Core Assessment Program

(Some sections have been omitted.)

The above examples show forms of assessment that produce numerical indicators and that also meet several criteria for authenticity. These procedures measure proficiencies under standard conditions that permit comparisons over time and between student groups; these assessments can also be used to assess mastery of specific curriculum content.

Exhibitions of Mastery

Exhibitions of mastery differ from the discrete competency tasks described above, but the difference is mainly a matter of degree. The exhibition is meant to reflect competence on challenging tasks that do not have a single clear solution; that require substantial time and student initiative; that require analysis, creativity, intellectual perseverance, and usually considerable integration of knowledge; and that result in products or performances that "stand alone" -- that is, they are likely to be aesthetically pleasing, relevant and informative, or have direct technical or social applications. In its ideal form, the exhibition demonstrates *mastery* of a particular topic or skill. Examples of exhibition projects include essays, speeches, public debates, research reports, computer demonstrations, video productions, engineering designs of different types, and artistic performances or products. An example from a school that uses exhibition as a form of assessment follows the description of the portfolio approach.

Portfolios

Portfolios furnish a broad, often longitudinal, portrait of individual performance on several dimensions. Like the portfolio assembled by a professional seeking a job, the student portfolio is a file or folder containing materials reflecting the student's experience and competence. Just as no single type of exhibition is best, no single approach excels in documenting personal characteristics and competence in a portfolio. However, portfolio contents do fall into three categories:

Samples of work, e.g., excerpts or complete pieces of writing; designs and blueprints; audio, video and photographic recordings of performances and projects.

Testimony, e.g., a student cumulative folder with letters of reference or taped testimony from authorities on the quality of student work or personal characteristics such as creativity, initiative, and dedication.

Summary indicators, e.g., a school transcript, test score results, or a personal resume including an employment record, evidence of membership in school clubs or academic events, a list of awards or any other distinctions, and other pertinent biographical data.

Portfolios may be highly open-ended, contain individualized statements or be more standardized in their content requirements. In the former case, they allow for greater individual expression; in the latter, standardization contributes to comparative evaluation. Portfolios may be comprehensive statements of a student's cumulative accomplishments in many areas over many years (and thus, operate like a resume or graduation transcript) or may be more focused statements documenting and illustrating growth in a particular domain such as mathematics or art (and thus be used pedagogically as a tool of diagnosis or assessment).[7]

A portfolio has legitimacy if its contents are validated by appropriate authorities, such as teachers, guidance counselors, or community representatives. For instance, signatures of an adult representative can help validate a student's successful participation in community service projects. In the case description below, the portfolio is validated by: 1) experts in the disciplines represented in the portfolio, 2) persons with more general knowledge of the candidate's character and abilities, and 3) a person who officially represents the school.

Exhibitions and Portfolios in Practice: The Case of Walden III High School

Walden III High School, Racine, Wisconsin, has a graduation requirement called the "Rite of Passage Experience," or R.O.P.E.[8] Prior to graduation, all seniors demonstrate mastery in 15 areas of knowledge and competence by completing a Portfolio, Project, and 15 Presentations before a R.O.P.E. committee consisting of staff (including the student's homeroom teacher), a student, and an outside adult.

The Portfolio. The Portfolio, developed during the first semester of the senior year, requires the graduating senior to reflect on his or her school experience, goals, and outside work. Among other components, it must include letters of recommendation for work, an autobiographical essay (which may include school records and other indicators of participation), a reading record (including a bibliography, annotated if desired, and two mini-book reports), and a series of written analytical reports on mass media, ethics, science and technology, and the arts.

The Project. Every graduating senior must write a library research paper analyzing an event or theme in American History. The student must be prepared to field questions on both the paper and an overview of American history during the second semester presentations.

Presentations. Each of the components of the Portfolio, plus the Project, must be presented orally and in writing to the R.O.P.E. committee. Additional oral presentations are also required, but written reports or new products are not required. Supporting documents or other forms of evidence may be used. Assessment of proficiency is based on multiple criteria including the organization and quality of the written products and the demonstration of knowledge and skills during the presentations in each of the prespecified areas.

During their first semester as seniors, students take a class in which they receive guidance on their R.O.P.E projects. They also must select their committee, develop and assemble the contents of their Portfolio, and write their Project paper. The Portfolio and the Project must be handed in to the R.O.P.E. instructor by the end of the first semester.

During the second semester, seniors present their Portfolio's contents, their research Project, and the other areas in which demonstrations of proficiency are required to their committees in an oral format. Presentations last about an hour and most require five to nine separate appearances before the committees.

The R.O.P.E. committees evaluate the portfolios, projects, and oral presentations. To earn a diploma, a student must receive a passing grade in at least 12 of the 15 required areas of the R.O.P.E. that also meet district requirements for proficiency in Math, Government, Reading, and English.

The R.O.P.E program reflects principles of authentic assessment in school restructuring. Instruction and assessment are intimately connected; students are expected to produce and demonstrate integrated forms of knowledge and competence; assessment criteria are clear and known in advance by students and staff; and the products of the students' intellectual efforts have value beyond the purpose of assessment. The fact that the R.O.P.E. program is a formal part of graduation and that assessment is done by teachers, parents, and students serving on committees creates a shared and institutionalized set of academic standards.

Walden III's approach is representative of a rich variety of interesting innovations in authentic assessment. Schools in the RE:Learning project in Delaware are making the "Exhibitions of Mastery" concept part of their restructuring agenda;[9] the Pittsburgh Public Schools use student portfolios to monitor and document academic growth and development and to individualize instruction. Other examples are described in Archbald and Newmann (in press), Burgess and Adams (1980), and in a forthcoming publication from the Congressional Office of Technology Assessment.

Examples of Authentic Assessment: Concluding Remarks

Authentic assessment of discrete competencies using exhibitions and portfolios covers much territory. Some approaches reflect more criteria of authenticity than others; some are more amenable to standardized administration than others; some require the student to draw on more integrated forms of conceptual knowledge than others; and some are more useful for diagnostic and instructional purposes while others are in the form of summative evaluations. All of them require students *to produce ideas*, *to integrate knowledge*, and all require *human judgment* in the application of *criterion-referenced judgments*. These basic principles are at the heart of authentic assessment.

The application of these ideas in a particular school setting will be more modest in scope than the full range of approaches described above. Which approaches are selected will depend

upon the purposes of assessment as viewed by school and district decision-makers. Implementation of these ideas must be preceded by a thorough assessment of needs and by careful deliberation over the purposes the assessments are intended to serve. Because many of these ideas are new, they cannot be implemented casually or hastily; effective training and leadership are essential.

AUTHENTIC ASSESSMENT AS AN AGENT OF CHANGE

Evidence of A Building Wave of Testing Reform

Research, theory, advocacy, and policy changes related to authentic assessment have increased dramatically in the last few years.[10] The majority of states as well as many districts and schools are making their testing more authentic through the use of portfolios of student work, essay examinations, writing samples, open-ended (paper-and-pencil) mathematics or science problems, hands-on exhibitions of problem-solving in mathematics and science, and other authentic measures. New York, for instance, recently assessed the performance of 200,000 fourth graders on a set of "hands on" science procedures, including how experiments are conducted.

The state of Vermont and the Pittsburgh Public Schools are using portfolios to assess student achievement and to produce diagnostic information for teachers and curriculum specialists. Maryland is implementing a revamped statewide testing program that includes a large number of authentic assessment tasks spanning the K-12 curriculum. The National Assessment of Educational Progress has piloted a large number of questions aimed at assessing conceptual understanding and problem solving, and current discussions indicate continued movement toward an expanded and more comprehensive national testing program, which will contain numerous authentic assessment tasks (e.g., Resnick, 1991, p. A12) U.S. Department of Education, 1990).

The Improvement Potential of Authentic Assessment

The early stages of most reforms are times of heady optimism and inspiring rhetoric. The enthusiasm for authentic assessment is typical in that much attention has been paid to reform potentials -- in particular, the potential of authentic assessment (1) to clarify and legitimate high

standards of academic achievement, (2) to create incentives to spend more instructional time on conceptual thinking and problem solving, and (3) to provide teachers, administrators, and policy-makers with useful pedagogical and management information. However, less attention has been paid to the limits of time and budgets and the daunting challenge of institutional change that authentic assessment will require. The following discussion treats the potentials of authentic implementation assessment.

Potential (1): Clarifying & Legitimating High Standards

Reform-minded administrators and policymakers see in authentic assessment the potential to broadcast a clear and authoritative message about desired educational outcomes. We can think of this potential as manifested at two levels: an institutional level and at the individual teacher/classroom level.

Institutionalizing authentic standards Currently, most schools, districts, parents, and policymakers take standardized test scores very seriously.[11] "Effectiveness," "quality," and "academic achievement" are often equated in educational discourse with high performance on standardized tests. Real estate agents show "good" schools and neighborhoods using test scores as the measure. Districts and states commonly proclaim goals of raising standardized achievement scores. The cover of a high-profile policy document in one state reads: "On a statewide average, educational achievement will equal that of the upper quartile of states within five years as indicated by commonly accepted criteria of achievement" (read, test scores).[12]

This kind of attention and seriousness may eventually be reserved for more authentic indicators of academic performance if they attain the legitimacy that formal testing confers. Resnick (1991, p. A12), commenting on exam systems in other countries, notes how requirements of academic achievement are part of a "public conversation" about standards:

> In many of these countries, the day after these tests, the questions show up in the newspapers and are discussed on television. There is a whole culture that shapes the public's understanding of what the standards are.

The objective is to institutionalize measures of performance that cannot be reduced to simple numbers and rankings. Too willingly policymakers and newspapers pander to a winner-loser mentality, sanctifying the test score and losing sight of what the numbers mean. The fact that authentic assessment involves multiple indicators and more descriptive measures of outcomes offers promise for re-focusing institutional attention on standards that really count.

Clarifying authentic achievement in the classroom "We teach what we test," has been oft-written and heard in authentic assessment discourse. More than any other mechanism, authentic assessment has the potential to clarify everyday classroom instructional goals and to maintain a focus on forms of academic competence that ultimately represent what it is we want students to know and be able to do. As Archbald and Newmann (1988) put it:

> Courses might well *begin* by presenting the "final exam" or other authentic assessment tasks. These tasks should cast a shadow back on the curriculum, serving as a constant target for study and practice, as when artists prepare for performances, when carpenters build a house, when athletes prepare for the big game. In short, tests, projects, and performances that demonstrate authentic academic achievement are valuable not only as assessment devices, but as guides to focus and to inspire teaching and learning.

In forcing one to think about and operationalize precise educational goals, there are few activities that compare with developing authentic assessment tasks and rating products and performances. Deciding what specific dimensions and criteria distinguish different levels of performance in scientific problem solving or spoken or written discourse is challenging professional work. It is work teachers should engage in more frequently, because it requires deep reflection on educational purposes and collaboration on academic standards. Multiple-choice testing is convenient, but pernicious in that it allows complacent inattention to fundamental questions about curriculum and instructional goals.

If authentic assessment conveys the expectation that problem-solving in science and mathematics, critical thinking in the humanities, and written and oral communication are essential academic outcomes, then the discrepancy between what schools teach and student performance will

become more visible and more difficult to reconcile. This may be a difficult or challenging realization, but a necessary one to marshall sorely needed resources and commitment to bring about curricular and instructional improvement.

Potential (2): Incentives for authentic instruction and tasks

Accountability adds incentive, making assessment instruments agents of curriculum change. The evidence is quite clear that accountability uses of standardized tests influence teachers' selection of teaching topics and methods. Cannell's (1988) landmark study uncovered the statistical oddity described as "nationally normed elementary achievement testing in America's public schools: How all 50 states are above average." Cannell found most school districts also report "above average" scores. This phenomenon results from a combination of two factors -- present results being compared to national norms derived 6 or 8 years ago, and teaching and curriculum focusing more directly on content domains tested, a process called instructional alignment. Classroom-focused studies have also described teachers deliberately tailoring their instruction toward topics covered on tests and toward the development of test-taking skills (Smith, 1991; Wilson & Corbett, 1989).

While curriculum-testing alignment and test preparation produces test score gains, what these gains mean is less clear. Gains may reflect genuine improvement in the kinds of educational outcomes assessed by multiple-choice tests.[13] But gains may also be artificial, a result of preparation drills and exercises that mirror test questions. This practice "pollutes" the measurement process and prevents valid inferences about what has caused the gains (Haladyna, Nolan, & Haas, 1991). Sadly, instruction aimed at producing quick gains in standardized test scores may depress the development of authentic academic knowledge and dispositions.

Can we apply the model of accountability-induced test score gains to authentic assessment policy? If children, teachers, and schools are held accountable for performance on a variety of authentic assessment instruments rather than exclusively on standardized test performance, will teachers focus instruction more on authentic academic tasks? Some evidence says, "Yes," but most is in the form of anecdotal accounts and observations rather than conclusive research and

evaluation. New York state's Regents examinations in history and global studies contain essay questions. Passing these exams (in addition to others in mathematics and science) confers a Regents diploma that enhances a student's college admissions prospects. Teachers have a stake in students' performance too -- not just because they want their students to do well, but also because administrators can review how well a given teacher's students perform on the Regents exams. These incentives insure that students practice essay-writing during the school year (Archbald, 1990). Since writing is an indisputably important academic task, the Regents exams -- at least the essay portion -- may well result in students spending more time writing than they otherwise would do. This outcome -- students writing more -- has also been claimed as a benefit from school and district-wide, holistic writing assessment programs. Because teachers and students know that writing will be tested, teachers assign and students do more writing, and more attention is paid to criteria of quality in writing.

Potential (3): More useful instructional/management information

Standardized test score information is produced relatively cheaply, but is also of limited usefulness to teachers and principals. There is some value in knowing how one's students or school compares to others in the district, state, or nation. However, the yardstick, a standardized, multiple-choice test and a measure that is relative, inevitably raises doubts and questions. A principal or teacher can rightfully ask: Are we teaching what the test covers? Are we spending too much time on what the test covers? Does time spent on hands-on learning, field trips, or classroom discussions detract from our test performance? How should a 5% or 10% change in percentile be interpreted? What does a "gain score" really mean? At what percentile can one be confident that the test performance genuinely signifies competence in reading, writing, science, or mathematics?

There is less "mystery" involved in interpreting scores or ratings on authentic assessments compared with standardized test scores. Authentic assessment involves (1) criterion referencing, meaning that a score has a fixed, specific referent: e.g., "on a scale of 1 to 5 on a writing test, my class averaged 3.7" (where a "3" and a "4" might be defined as written in Figure 3); and (2) more

authentic tasks in the assessment of performance. In authentic assessment, the test is closer to the "real thing;" measurement and interpretation are more direct. One makes students write in order to assess writing competence or perform a science experiment to assess knowledge of scientific procedures. In contrast, when interpreting standardized test scores, one must infer levels of academic competence from performance on multiple-choice questions.

Results of assessments on authentic tasks, then, yield relatively straightforward information about what students can and cannot do and have and have not learned. Assessment and instruction are closely connected; each has direct implications for the other. As one researcher concluded after case studies of six different authentic assessment programs across the nation: "it was difficult to keep respondents focused on the assessment. Teachers and administrators so often experienced [their assessment systems] as instructional programs that we may have received answers about the program when we thought we were asking about the assessment..." (Mitchell & Stempel, forthcoming).

The potentials of authentic assessment outlined above form a model of curriculum and instructional improvement in which assessment plays a key role. The model relies essentially on three factors to initiate and guide improvement: goal clarification, accountability incentives, and the administrative and instructional usefulness of performance information. Like any model, it must ultimately come to grips with the realities of politics and practice. While the intent of this paper is mainly to describe principles and examples of authentic assessment and to discuss their potential as instruments of reform, it is necessary to acknowledge potential problems in the translation of principles into practice.

Difficult Realities of Implementation and Reform

Every idea's theoretical potential for reform faces many reasons it may not succeed. Limited budgets, political forces, and technical constraints will shape the implementation of authentic assessment, making practice inevitably fall short of ideals. Prudent policies and

administration can be the difference between ideas that fail and those that push the limits of their potential, and in some instances become ideals.

Authentic assessment has the potential to shift our preoccupation with standardized test outcomes toward a more enlightened conception of organizational quality and individual academic competence. However, the desire for simple numerical rankings of schools and students will not diminish without considerable political effort and public education. Standardized test manufacturers have an economic interest in sustaining use of standardized testing, legislators gain political mileage from advocating testing to insure accountability to taxpayers, and newspapers and their audiences like to identify winning and losing schools in the arena of public education. School people intending to embark on innovations in authentic assessment will do well to prepare a defense against skeptical questioners, many of whom will persist in seeing standardized test scores as the ultimate benchmark against which authentic assessment measures should be validated.

Authentic assessment is costly and time-consuming. Time and money are consumed *in staff development* -- few teachers and principals are familiar with the concepts and procedures of authentic assessment; *in test development* -- unless a district or school uses a ready-made set of authentic assessments (some are available), it will need to expend significant resources in developing its own assessment approaches uniquely tailored to its own curricular needs; and *in administering authentic tests* -- "manual" grading, whether by observing and evaluating performances or collecting and evaluating products, is a time-consuming process. Because authentic assessment is an investment in multiple purposes -- organizational and staff development as well as curricular and instructional reform -- cost-benefit analyses are more complex.

People will compare the time and resources expended in authentic assessment to standardized testing costs, and some will dismiss authentic assessment out of hand. This is a mistake in perspective. With standardized tests, schools and districts pay for test score information with significant limits of authenticity and usefulness; at the same time, there is little if any "return" on the investment. Some argue, indeed, that "losses" are incurred in that educators abdicate responsibility for defining and setting standards, teachers are removed from the assessment

process outside of their own classrooms, the public may well receive invalid (i.e., polluted) performance information, and genuine accountability has not been achieved.

In contrast, the return on the investment in authentic assessment is described in the potentials above. Teachers become partial owners of a larger process of organizational and academic assessment. Teachers learn, and grow professionally, from developing, participating in, and observing a more complex and authentic assessment process. And teachers gain detailed and useful information about the effectiveness of their curriculum and instruction, and needs for improvement.

Perhaps most important, authentic assessment can contribute to a more enlightened conception of competence. The *meaning* of academic achievement or intellectual competence should evoke much more than recall of information and skill in doing simple tasks or reproducing formulas or pat answers. We need as a society to reach the point where achievement and competence are not defined by minimums, but instead evoke conceptions of persons performing confidently and competently to analyze complex issues and solve real problems. Getting to that point can only be accomplished one step at a time. Making testing more authentic is a key step in the process.

REFERENCES

American Association for the Advancement of Science. (1989). Science for all Americans. [Project 2061 report on Literacy Goals in Science, Mathematics, and Technology.] Washington, DC: Author.

Anastasi, A. (1988). Psychological testing (6th Edition). NY: Macmillan.

Archbald, D. (1990, April). Curriculum control and teacher autonomy. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Archbald, D. & Newmann, F. M. (in press). Approaches to assessing organizational and individual academic performance. In H. Berlak & F. M. Newmann (Eds.), Toward a new science of educational testing and assessment. Albany, NY: State University of New York Press.

Archbald, D. & Newmann, F. (1988). Beyond standardized testing: Assessing authentic academic achievement in the secondary school. Reston, VA: National Association of Secondary School Principals.

Archbald, D., & Porter, A. (1991). A retrospective and an analysis of the roles of mandated testing in education reform. Washington, DC: Congressional Office of Technology Assessment.

Baxter, G., Shavelson, R., Pine, J., & Yure, J. (1991, April). Hands-on performance assessments and their pencil and paper surrogates. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Burgess & Adams. (1980) Outcomes of education. London: Macmillan.

Council of Chief State School Officers. (1990). Performance assessment in the states. Prepared by the Council of Chief State School Officers for the Assessment Committee of the Secretary's Commission on Achieving Necessary Skills. Pelavin Associates, Washington, DC.

Cannell, J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above average. Educational Measurement: Issues and practice, 7(2), 5-9.

Committee for the Assessment of Experiential Learning. (1975). A guide for assessing prior experience through portfolios (Working Paper No. 6). Princeton, NJ: Educational Testing Service.

Forrest, A. (1975). A student handbook on preparing a portfolio for the assessment of prior experiential learning. CAEL Working Paper No.7. Princeton, NJ: ETS.

Congressional Office of Technology Assessment. Performance assessment: Methods and characteristics. Washington, DC: Author.

Forrest, A., & Steele, J. M. (1982). Defining and measuring general education knowledge and skills (Tech. Rep. No. 1976-81). Iowa City, IA: College Outcome Measures Project, The American College Testing Program.

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem solving. Applied Psychological Measurement, 2, 1-24.

Gorman, T. (1986). The Framework for the assessment of language, published by NFER-NELSON, Darville House, 2 Oxford Road East, Windsor, Berkshire SL41DF, England.

Gorman, T., White, J., & Brooks, G. (1982). Language performance in schools. Report on the 1982 secondary survey from the Language Monitoring team at the National Foundation for Educational Research in England and Wales to the Department of Education and Science, the Welsh Office and the Department of Education for Northern Ireland.

Haladyna, T., Nolan, S., & Haas, N. (1991, June-July). Raising standardized achievement test scores and the origins of test score pollution. Educational Researcher, 20(5), 2-8.

Jencks, C., Bartlett, S., Corcoran, M., Crouse, J., Eaglesfield, D., Jackson, G., McClelland, K., Mueser, P., Olneck, M., Schwartz, J., Ward, S., & Williams, J. (1979). Who gets ahead? New York: Basic Books.

Keefe, J. W. (Ed.) (1988). Profiling and utilizing learning style. Reston, VA: National Association of Secondary School Principals.

Linn, R. L. (1982). Ability testing: Individual differences, prediction, and differential prediction. In A. K. Wigdor & W. R. Garner (Eds), Ability testing: Uses, consequences, and controversies, Part II (pp. 335-388). Washington, DC: National Academy Press.

Linn, R. & Baker, E. (n.d.) Demands for performance-based assessment and criteria for their validation. UCLA Center for the Study of Evaluation, University of California-Los Angeles.

Maclure, M., & Hargreaves, M. (1986). Speaking and listening: Assessment at Age 11. Published by NFER-NELSON, Darville House, 2 Oxford Road East, Windsor, Berkshire SL41DF, England.

Mehrens, W., & Lehmann, I. (1987). Using standardized tests in education. New York, NY: Longman

Mentkowski, M., & Doherty, A. (1982). Validating assessment techniques in an outcome-centered liberal arts curriculum: Integrated competence seminar. Milwaukee, WI: Alverno College.

Mitchell, R. & Stempel, A. (forthcoming). Six case studies of performance assessment. Washington, DC: Office of Technology Assessment.

National Commission on Social Studies in the Schools.(1989). Charting a course: Social Studies for the 21st century. Report of the Curriculum Task Force on the National Commission on Social Studies in the Schools.

National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

National Research Council. (1989). Everybody counts. Washington, DC: National Academy Press.

Neill, M. & Medina, N. (1989). Standardized testing: harmful to educational health. Phi Delta Kappan, 70(9), 688-697.

Newmann, F. & Archbald, D. (in press). The nature of authentic academic achievement. In H. Berlak & F. M. Newmann (Eds), Toward a new science of educational testing and assessment. Albany, NY: State University of New York Press.

Owen, D. (1985). None of the above: Behind the myth of scholastic aptitude. Boston, MA: Houghton Mifflin.

Resnick, L. (1991, May 20). Tests in other nations studied for hints to US. New York Times, A12.

Rothman, R. (1989, March 22). District ties goals to test scores. Education Week, p.1.

Sizer, (1984). Horace's compromise. Boston, MA: Houghton Mifflin.

Smith, M. (1991, June-July). Put to the test: The effects of external testing on teachers. Education Researcher, 20(5), 8-11.

Tindal, G. & Parker, R. (1989). Development of written retell as a curriculum-based measure in secondary programs. School Psychology Review, 18(3), 328-343.

U.S. Department of Education. (1990). National goals for education. Washington, DC: Author

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. Journal of Educational Measurement, 17(1), 11-29.

Whimby, A. (1985). You don't need a special "reasoning" test to implement and evaluate reasoning training. Educational Leadership, 43(2), 37-39.

Wilson, B. & Corbett, H. (1989). Two state minimum competency testing programs and their effects on curriculum and instruction. Philadelphia, PA: Research for Better Schools.

ENDNOTES

1. See for example articles by Owen, D. (1985) and Neill & Medina (1989).

2. See Jencks et al. 1972:60; Whimbey 1985.

3. For a general overview of validity issues in performance assessment, see Linn & Baker (n.d.). Mehrens & Lehmann (1987) is an excellent resource on the design, measurement properties, and purposes of standardized tests.

4. It bears noting that corporate America's desire to compete better with corporate Japan has brought re-evaluation of the efficiency of individualistic assembly-line models of production. Many corporations have emulated Japan's use of "quality teams" and "quality circles" composed of laborers responsible to each other and also responsible for larger units of production. Both workers and students benefit from collaboration on an integrated product as opposed to working individually on fragments whose connections to larger wholes are less clear.

5. The American College Testing corporation, through the College Outcome Measures Project (Forrest & Steele, 1982), provides a useful framework for categorizing educational outcomes. It has developed a set of instruments for the assessment of proficiencies in applying specific facts and concepts in work, family, and community roles.

6. For more information about the APU and a complete description of the assessment tasks described in "The Assessment of Speech," see: Gorman, T., White, J., and Brooks, G. (1982); Gorman, T. (1986); and Maclure, M. and Hargreaves, M. (1986).

7. For more on portfolios, see Committee for the Assessment of Experiential Learning (1975) and Forrest (1975).

8. This example is drawn from Archbald and Newmann (1988).

9. For an eloquent rationale for restructuring and the use of exhibitions of mastery for assessment, see Sizer (1984).

10. The National Association of Secondary School Principals Learning Style Profile provides diagnostic information on various performance measures of cognitive skills; see Keefe (1988). For information on performance assessment policy developments, see CCSSO (1990).

11. An illuminating case in point is described in Rothman (1989, March 22).

12. Student Performance Standards of Excellence in Mathematics, Science, Social Studies, and English, Tallahassee, FL: Department of Education, 1984.

13. For discussion of these issues see Archbald & Porter (1991).